Is Large-Scale *Ab Initio* Hartree-Fock Calculation Chemically Accurate? Toward Improved Calculation of Biological Molecule Properties

HAJIME TAKASHIMA,[1] KUNIHIRO KITAMURA,[1] KAZUTOSHI TANABE,[2] UMPEI NAGASHIMA[2,3]

[1] *Medicinal Research Laboratories, Taisho Pharmaceutical Co., Ltd., 1-403 Yoshino-cho, Ohmiya, Saitama 330, Japan*
[2] *National Institute of Materials and Chemical Reaction, Ibaraki, Japan*
[3] *Department of Information Sciences, Faculty of Science, Ochanomizu University, Tokyo, Japan*

Received 14 April 1998; accepted 31 October 1998

ABSTRACT: Numerical errors in total energy values in large-scale Hartree–Fock calculations are discussed. To obtain total energy values within chemical accuracy, 0.01 kcal/mol, stricter numerical accuracy is required as basis size increases. In molecules with 10,000 basis sizes, such as proteins, numerical accuracy for total energy values must be retained to at least 11 digits (i.e., to the order of 1.0D-10) to keep accumulation of numerical errors less than the chemical accuracy (0.01 kcal/mol). With this criterion, we examined the sensitivity analysis of numerical accuracy in Hartree–Fock calculation by uniformly replacing the last bit of the mantissa part of a double-precision real number by zero in the Fock matrix construction step, the total energy calculation step, and the Fock matrix diagonalization step. Using a partial summation technique in the Fock matrix generation step, the numerical error for total energy value of molecules with basis size greater than 10,000 was within chemical accuracy (0.01 kcal/mol), whereas with the conventional method the numerical error with several thousand basis sets was larger than chemical accuracy. Computation of *one* Fock matrix element with parallel machines can

Correspondence to: H. Takashima
Contract/grant sponsor: New Energy and Industrial Technology Development Organization
Contract/grant sponsor: Institute of Research and Innovation
Contract/grant sponsor: Science and Technology Agency
Contract/grant sponsor: Ministry of Education, Science and Culture

CCC 0192-8651 / 99 / 040443-12

include the partial summation technique automatically, so that parallel calculation yields not only high-performance computing but also more precise numerical solutions than the conventional sequential algorithm. We also found that the numerical error of the Householder-QR diagonalization routine is equal to or less than chemical accuracy, even with a matrix size of 10,000. J Comput Chem 20: 443–454, 1999

Keywords: chemical accuracy; numerical errors; Fock matrix generation and diagonalization; sensitivity analysis; partial summation technique

Introduction

Throughout science and technology, large-scale calculations not previously performed are being carried out as a result of the surprising development and proliferation of computing machines, especially high-performance workstations, and development of faster algorithms. Quantum chemistry is no exception. Many scientists, including quantum chemists, now execute large-scale molecular orbital calculations[1–4] with basis sizes > 1000 using parallel processors such as workstation clusters and supercomputers. The scaling properties of the Hartree–Fock method have been evaluated on carbon–hydrogen model systems.[5] Scalable and efficient algorithms suitable to parallel systems have been proposed[6, 7] to perform calculations for very large molecules using parallel processors. A speed-up has also been achieved with the development of some faster algorithms.[4, 8, 9] Because such large-scale numerical simulation is indispensable for understanding the mechanisms of chemical phenomena and designing new medicines and materials, the importance of large-scale molecular orbital calculations will increase in the future.

Large-scale calculations include so many operations that the reliability of the solution obtained depends greatly on the number of the mantissa bit used in the computation. At present, almost all *ab initio* molecular orbital programs use a double-precision real number, as do many other scientific calculation programs. However, there is no guarantee that sufficiently accurate results will be obtained in large-scale calculations with the present double-precision calculation. We are now developing a special-purpose machine[10, 11] that computes electron repulsion integrals and generates Fock matrix elements in *ab initio* molecular orbital calculations with superhigh speed. This machine targets large and nonsymmetric molecules, like proteins, and therefore we wish to determine whether the double-precision real number is sufficiently accurate for calculations performed for such extremely large molecules. In addition, design of a custom processor for the special-purpose machine requires evaluation of numerical errors in calculations.

Why do numerical calculation errors in large-scale molecular orbital calculations need to be considered? Whenever energy values are discussed using molecular orbital calculations, we compare the difference in total energy values between one state and another. These values are always quite large and almost the same, so several-digit cancellation occurs on subtraction of these values. For example, consider evaluation of the stability of the glycine molecule in two states, neutral type and twitterion type, in a vacuum. In a Hartree–Fock calculation with 6-31G basis sets, the total energy value of the neutral type is -177389.3065 kcal/mol, whereas that of the twitterion type is -177320.5292 kcal/mol. We therefore judge the neutral type to be 68.78 kcal/mol more stable than the twitterion type. In this case, four-digit cancellation occurs. Other properties, such as atomic charges and dipole moments, require relatively little numerical accuracy of only several digits, because they are usually evaluated not by subtracted values but by themselves, and such several-digit cancellation does not occur. Thus, total energy values require the greatest numerical accuracy of all the properties that can be calculated in molecular orbital calculations.

Energy difference is usually on the order of several kilocalories per mole or less, and we must obtain total energy values with an absolute accuracy on the order of several kilocalories per mole for both large and small molecules. However, total energy values for large molecules are much larger than those for small molecules. Thus, total energy values for large molecules require relatively greater numerical accuracy than those for small molecules, because larger cancellations can be expected on subtraction.

On the other hand, most molecular orbital calculation programs use a double-precision real number, containing 53 bits as a mantissa. Its effective relative accuracy corresponds to the machine epsilon[12] and is about 2.22D-16. Large numbers of operations therefore accumulate large numerical errors.

Therefore, in the *ab initio* molecular orbital calculations of relatively large molecules, like proteins, greater numerical accuracy for total energy values is required although numerical errors are accumulated to a greater extent. For this reason, calculated total energy values should contain numerical errors greater than the lowest value with chemical meaning.

The organization of this article is as follows: First, we discuss the relationship between molecular size and the required relative accuracy of its total energy value. Next, using the results obtained with numerical experiments, we clarify that the present Fock matrix construction algorithm cannot yield sufficient chemical accuracy; that is, the estimated numerical error for total energy value is larger than the required chemical accuracy if the number of basis set is larger than several thousand. We also show how to solve this problem. Finally, we discuss the numerical errors in the diagonalization step.

Total Energy Values and Required Accuracy

In this section, we clarify the relationship between the number of basis functions and required numerical accuracy relative to a total energy value. In discussions of the stability of molecules, we compare the total energy values of each system. The energy difference is usually referred to the experimental values, for which heat fluctuation is on the order of 0.1 kcal/mol. To obtain a reliable calculated energy value, its numerical error must be at least one digit smaller than the order of heat fluctuation. We thus regard 0.01 kcal/mol as the required absolute accuracy for energy values. This is the lowest energy value that has chemical meaning. Using this value, we can estimate the required relative accuracy for total energy values with the following equation:

$$(\text{relative accuracy}) = \frac{0.01}{c \times (\text{total energy value [a.u.]})}$$

Here, c is the unit conversion coefficient from atomic unit to kilocalories per mole and is 627.5096. Because total energy value increases infinitely with molecule size, we fixed the upper limit at 10,000 basis functions. This corresponds nearly to the number of basis functions in biological molecules, like proteins, which we target in development of our special-purpose machine.

Total energy values and required relative accuracy for some large molecules are listed in Table I. The columns from the left give the molecule name, the kind of basis function, number of basis functions used, number of basis functions if a 6-31G basis is used, total energy values (a.u.), required relative accuracy, and reference citations. Molecules are sorted in the order of the number of 6-31G-converted basis functions. Figure 1 shows a graph plotted with the number of basis functions and required relative accuracy. The asterisks represent molecules including heavy atoms below the third period. Six compounds were calculated for this study and geometries of all peptides were assumed to be α-helix, which is the most compact structure. Calculations were executed on an SGI-Cray R10000, and the GAUSSIAN-94 program[13] was used.

Results are shown in Table I and Figure 1. First, a good linear relationship was obtained between the number of basis functions and relative numerical accuracy for total energy values when both were plotted with a logarithmic scale. Stricter relative numerical accuracy is required in proportion to the number of basis functions. For example, the required relative accuracy is around 3.0D-8 for 100 basis functions and around 3.5D-9 for 1000 basis functions. In the extrapolations along the regression line fitting to the crossmarks, it can be seen that the required accuracy with 10,000 basis functions is around 4.0D-10. Second, the linear relationship obtained applies only to molecules consisting mainly of hydrogen elements in the second period. Notably, if the molecule has many heavy atoms below the third period, the required accuracy is much stricter than that for molecules with only light atoms but almost the same basis size.

We have thus shown here that the requirement for relative numerical accuracy of total energy values increases proportionately to the number of basis functions. To obtain a reliable total energy value, with a numerical error less than chemical accuracy of 0.01 kcal/mol, the relative numerical error must be less than approximately 1.0D-8 for

TABLE I.
Total Energy Values and Required Relative Accuracy for Some Large Molecules.

Molecule	Basis function	Number of basis	6-31G converted	Total energy[a] (a.u.)	Required accuracy	Refs.
$ZnCl_2$	STO-3G	36	45	−2666.57	5.97D-09	—[b]
Gly	6-31G	55	55	−282.69	5.64D-08	—[b]
ZnI_2	STO-3G	52	61	−15458.74	1.03D-09	—[b]
Gly–Ala	6-31G	110	110	−528.18	3.03D-08	—
$(H_2SiO)_5$	6-31G*	190	130	−1825.2	8.73D-09	14
Gly–Ala–Gln	6-31G	207	207	−980.5	1.63D-08	—[b]
Gly–Ala–Gln–Met–Tyr	6-31G	427	427	−2251.42	7.08D-09	—[b]
Kekulene	6-31G**	792	480	−1831.86	8.70D-09	15
C_{60}	4-31G	540	540	−2268.52	7.02D-09	16
DMPC (lipid molecule)	6-31G*	838	562	−2399.14	6.64D-09	17
Chlorophyll dimer	DZ	1100	986	−4491.15	3.55D-09	18
EDP	3-21G	1461	1461	−8154.72	1.95D-09	4
Chlorophyll tetramer	DZ	2200	1972	−8988.92	1.77D-09	18
CRD	3-21G	3237	3237	−16737.91	9.52D-10	4
P53	3-21G	3836	3836	−17115.3	9.31D-10	4

[a] 1 a.u. = 627.5096 kcal / mol.
[b] Calculations for this molecule performed in the present study.

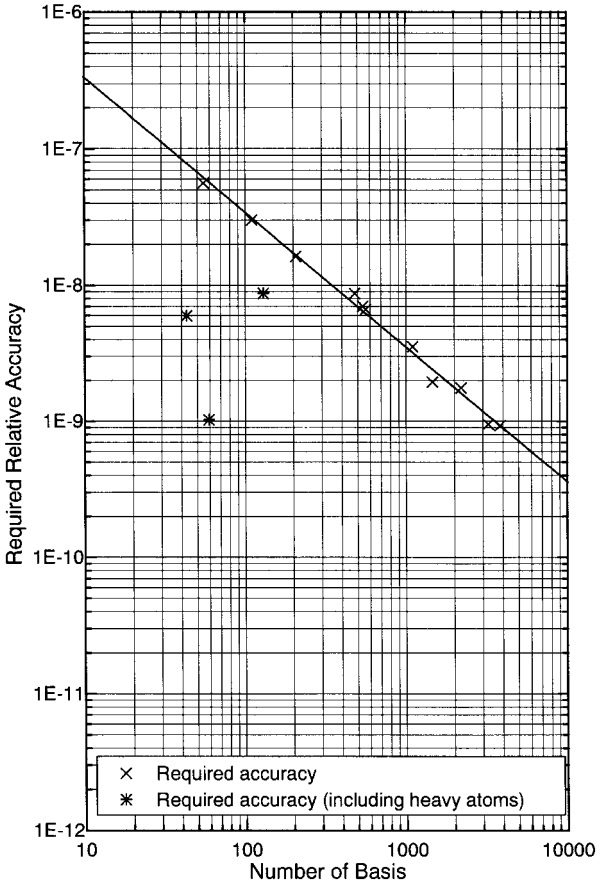

FIGURE 1. Required relative accuracy for total energy values.

300 basis functions, 1.0D-9 for 3000 basis functions, and about 3.0D-10 for 10,000 basis functions.

Numerical Errors

Almost all *ab initio* molecular orbital program packages now use a double-precision real number calculation, and we can assume that the Fock matrix generation, total energy calculation, and Fock matrix diagonalization steps cause accumulation of the largest numerical errors in the Hartree–Fock calculation. We therefore examined the numerical errors caused by these steps.

FOCK MATRIX GENERATION AND ENERGY CALCULATION STEPS

The equation for calculation of Fock matrix elements, F_{ij}, is:

$$F_{ij} = H_{ij} + \sum_{k=1}^{N} \sum_{l=1}^{N} P_{kl}\{(ij, kl) - \tfrac{1}{2}(ik, jl)\} \quad (1)$$

where H_{ij}, P_{ij}, N, and (ij, kl) denote the one-electron integral matrix element, the density matrix element, the number of basis functions, and the two-electron integral value, respectively. The Fock matrix generation algorithm *faithfully* correspond-

ing to eq. (1) is as follows:

C compute F_{ij}
DO k = 1, N
DO l = 1, N
calculate $(ij, kl), (ik, jl)$

$$F_{ij} = F_{ij} + P_{kl} * \left\{ (ij, kl) - \frac{1}{2} * (ik, jl) \right\}$$

END DO
END DO
$F_{ij} = F_{ij} + H_{ij}$

The second term of the Fock matrix computation equation in do loops is the cause of problems with numerical accuracy, because the addition times to calculate one Fock matrix element are the square of N. In a large-scale calculation, the numerical error is thus rapidly accumulated. The model algorithm just shown is a Fock-matrix-index-driven algorithm, but the algorithms actually used in almost all molecular orbital program packages are integral-driven forms that fully use the symmetry of the two-electron integral:

$$(ij, kl) = (ij, lk) = (ji, kl) = (ji, lk) = (kl, ij) = (kl, ji) = (lk, ij) = (lk, ji)$$

However, in regard to addition times, the Fock-matrix-index-driven algorithm just described is essentially the same as the integral-driven algorithm. So, we consider here the Fock-matrix-index-driven algorithm to make subsequent discussion simpler. We also note that the neglect of the smaller integral values causes the decrease in addition times to Fock matrix elements.

Using Fock matrix elements, the total energy value, E, can be represented by:

$$E = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} P_{ij}(H_{ij} + F_{ij}) \qquad (2)$$

The problem is the same as the eq. (1). The addition times to calculate the total energy value are the square of N. This case requires more accurate results, because all Fock matrix elements must be summed and no cutoff occurs.

To observe the numerical errors caused by these computing steps, we uniformly replaced by zero to the last bit in the mantissa part of the input data (two-electron integrals and density matrix elements), and output data (Fock matrix elements), in the Fock matrix construction step. We made the same replacements in the total energy calculation step. Alternatively, these operations can be regarded as error contamination in the last bit of the double-precision number. We then compared the total energy values obtained with those calculated in the normal double-precision type. We used the GAUSSIAN-94 program[13] and modified it to operate the bit of the double precision real number. Because it is quite difficult to perform calculations for molecules with more than 1000 basis functions with the computing machines presently available, we performed calculation for the peptides glycine, Gly–Ala (GA), Gly–Ala–Gln (GAQ), and Gly–Ala–Gln–Met–Tyr (GAQMY). With 6-31G basis sets, the minimum number of basis functions was 55 for the glycine molecule and the maximum was 427 for GAQMY. To explore the effect of type of basis set, we also used STO-3G and 6-31G** basis sets in addition to 6-31G basis sets. We did not perform calculations of GAQMY molecules with the 6-31G** basis set because of the enormous calculation time required. It may be appropriate to use the stricter threshold values to keep all integrals and to calculate more precisely. However, what we intend to investigate now is not the error from calculation techniques (screening effect, etc.) but the actual accumulation of numerical errors (round-off error). They are related but not the same problem. So we used the keyword scf = tight and default threshold values for calculations. The relative numerical error is defined as:

$$\text{(Relative numerical error)} = \left| \frac{(\text{Num53} - \text{Num52})}{\text{Num53}} \right| \qquad (3)$$

where Num53 denotes the numerical solution calculated in the default fashion and Num52 is the numerical solution calculated using the bit mask operation. Considering the larger basis sizes, we estimated numerical errors for regression lines and extrapolation.

The relationship between the number of basis functions and numerical errors for total energy values with the 6-31G basis set is shown in Figure 2a. We also performed calculations for these peptides with STO-3G and 6-31G** basis sets to examine the influence of type of basis set. The results are shown in Figure 2b and c, respectively. Table II shows data summarized from Figure 2. A fair degree of linear dependency was obtained by plotting on a logarithm scale. The linear line in the upper area is the same line as exhibited in Figure 1 and shows the numerical accuracy required to obtain chemical accuracy. The shadow ranges in

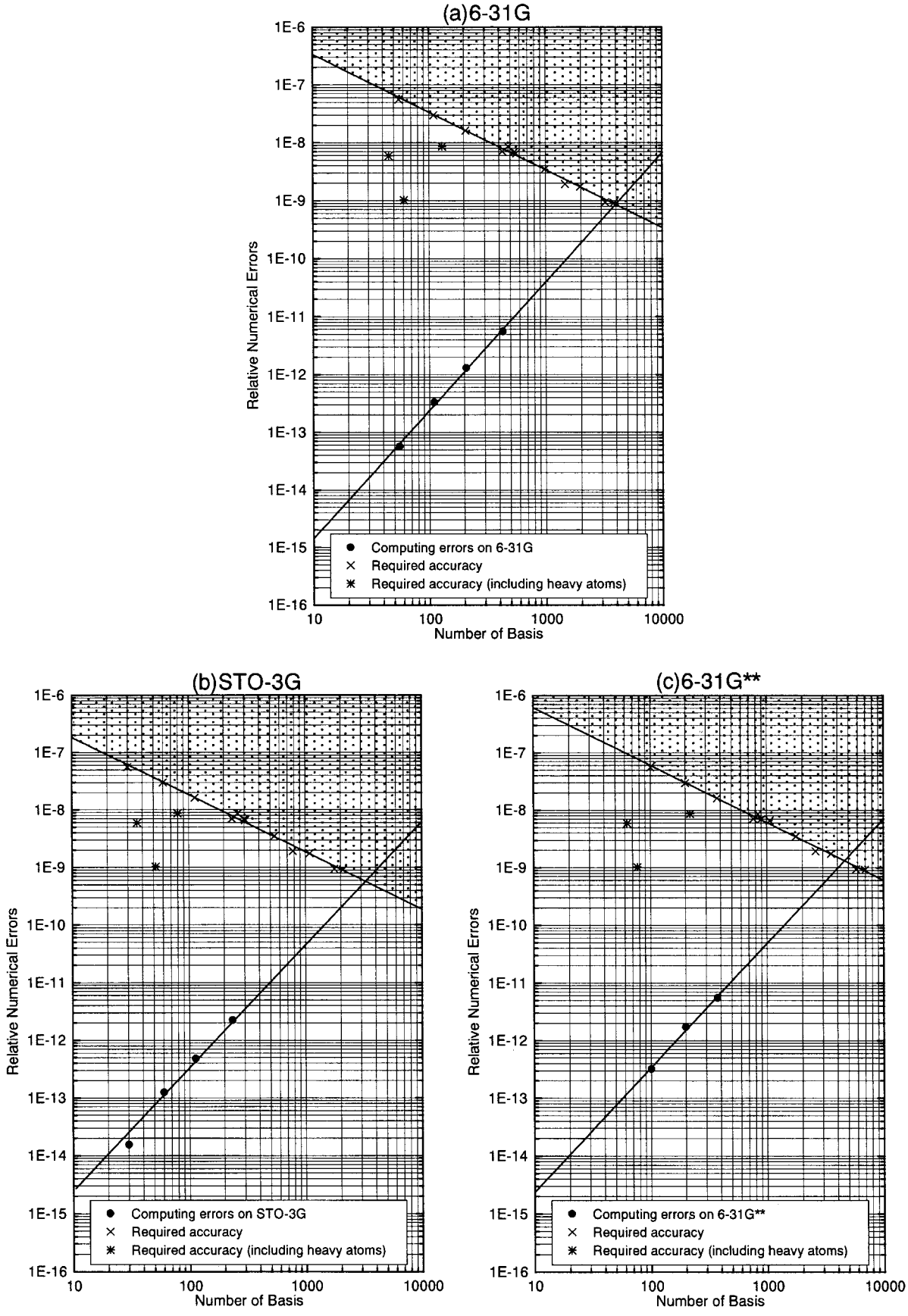

FIGURE 2. Relative numerical errors and required accuracy of total energy.

TABLE II.
Relative Numerical Errors for Total Energies.

Molecules	(a) 6-31G		(b) STO-3G		(c) 6-31G**	
	Basis[a]	Errors	Basis[a]	Errors	Basis[a]	Errors
Gly	55	5.73E-14	30	1.55E-14	100	3.22E-13
GA	110	3.32E-13	60	1.25E-13	200	1.72E-12
GAQ	207	1.30E-12	113	4.81E-13	375	5.53E-12
GAQMY	427	5.58E-12	235	2.25E-12	769	—

[a] Indicates the number of basis functions.

Figure 2 indicate that the relative numerical error included in the total energy value of the given number of basis sets is greater than the chemical accuracy of 0.01 kcal/mol. That is, when the line that shows the numerical error is in the shadow range, the total energy value obtained for the indicated number of basis functions is not reliable and may contain large numerical errors. For example, in the calculation with 200 basis functions (which corresponds to about 35 atoms), the relative numerical error of the total energy value is about 1.0D-12, whereas the maximum acceptable relative error for keeping chemical accuracy is about 1.0D-8. Therefore, this calculation is sufficiently reliable. However, numerical errors rapidly become larger as the basis size increases. For 4000 basis functions, the estimated numerical error is 9.0D-10 and the required accuracy is almost the same, so that the total energy value with > 4000 basis functions contains a large numerical error and is not sufficiently reliable.

Accumulation of relative numerical errors in floating-point arithmetic is estimated to be:

$$(1 + 2^{-m})^n \approx 1 + n \times 2^{-m} \qquad (4)$$

where m and n denote the number of mantissa bits and number of operations, respectively. In this case, m is 53 and n corresponds to the square of N, where N is the number of basis functions. When we calculate Fock matrix elements and total energy value with 10,000 basis sets the relative error is expected to be:

$$(1 + 2^{-53})^{10,000 \times 10,000} - 1 \approx 10^8 \times 2^{-53} \approx 1.11 \times 10^{-8} \qquad (5)$$

The estimated numerical error in total energy value is about 6.0D-9 in Figure 2a, and when the number of basis functions becomes tenfold larger, the relative numerical error is about 100 times larger. This indicates that error estimation in eqs. (4) and (5) is reasonable, and that large numerical errors occur in the Fock matrix construction step and the total energy calculation step.

In Figure 2b and c, if the number of basis functions is the same as that in Figure 2a (i.e., 6-31G basis sets) then the numerical errors obtained for total energy values have almost the same values and tendency. This indicates that numerical errors do not depend on the type of basis set, but are mainly affected by the number of basis functions. The result for the glycine molecule with 30 basis functions in Figure 2b deviated significantly from the line, perhaps because the Fock matrix construction steps and energy calculation steps have such small numbers of operations that numerical errors occurring there are hidden in those from other parts of the calculation.

On the other hand, the relative accuracy required depends to some extent on the type of basis set. The reason is that the molecule size described in each type of basis set differs considerably even if the basis size is the same. For example, 1000 basis functions corresponds to the chlorophyll dimer with a 6-31G** basis set, and also corresponds to the chlorophyll tetramer with a STO-3G basis set, so that the required accuracy is 6.6D-9 for 6-31G** and 2.0D-9 for STO-3G. Thus, if the number of basis functions is the same, calculation results with small basis sets (like STO-3G) require more precision than those with widespread basis sets (like 6-31G* or more extended basis sets). The basis size for which numerical errors are larger than the required accuracy is (as shown in Fig. 2) about 5000 for 6-31G** and about 2000 for STO-3G.

We must emphasize one significant point; that is, the lines showing the required accuracy in Figure 2 are applied to molecules consisting of only light atoms such as hydrogen and second period elements. The required accuracy for molecules containing many heavier atoms, such as those indicated by asterisks in Figure 1, increases because

the total energy values for such molecules are greater than those of molecules with only light atoms with the same basis size. Therefore, in molecules with many heavier atoms, the upper limit of number of basis functions to obtain sufficient chemical accuracy should be, for 6-31G, much lower than 4000.

IMPROVEMENT OF NUMERICAL ACCURACY

We showed that the total energy values have insufficient accuracy and contain large numerical errors for large-scale calculations, such as those with 10,000 basis functions in the previous subsection. Therefore, how can we decrease the number of numerical errors and execute numerical calculations with sufficient precision?

Large numerical errors are caused by many addition times to the same places. The partial summation technique is useful for decreasing the incidence of such errors. The model algorithm for Fock matrix construction using a partial summation technique is as follows:

C Compute the Fock matrix by
C partial summation
DO k = 1, N
$G_{ij} = 0.0$ D00
DO l = 1, N
calculate (ij, kl), (ik, jl)

$$G_{ij} = G_{ij} + P_{kl} * \left\{ (ij, kl) - \frac{1}{2} * (ik, jl) \right\}$$

END DO
$F_{ij} = F_{ij} + G_{ij}$
END DO
$F_{ij} = F_{ij} + H_{ij}$

where G_{ij} is the temporary array for partial summation. The total energy calculation step can be also written using partial summation as above.

Using this partial summation algorithm, we estimated the numerical errors for total energy values by the same bit masking operation as used in the previous subsection. We performed these calculations only with the 6-31G basis set, and the results are shown in Figure 3 and Table III. The results obtained include four points, and these values deviate so much from the regression line that the line may not be appropriate. However, the numerical errors are greatly decreased by partial summing. For the basis size of 200, the numerical error with conventional calculation is about 1.2D-12, whereas with partial summation it is only about 1.0D-14. Partial summation thus yields almost 100-fold greater precision. For basis size of 10,000, the expected numerical error is about 7.0D-9 with conventional calculation, much greater than the required accuracy of 3.0D-10. On the other hand, numerical accuracy using the partial summation technique is improved by about 10,000 times, and the value is 2.0D-12. This degree of accuracy is sufficient. Using the partial summation technique, numerical accuracy can be greatly improved and it is possible to obtain results with-

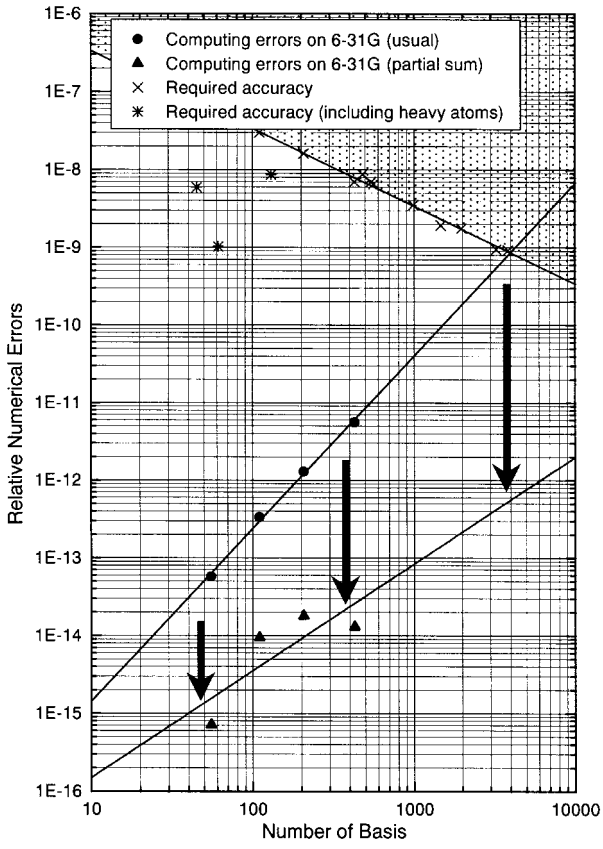

FIGURE 3. Relative numerical errors for total energy values determined by partial summation.

TABLE III.
Relative Numerical Errors for Total Energies Calculated by Partial Summation.

		6-31G	
Molecules	Basis[a]	Conventional	Partial
Gly	55	5.73E-14	7.10E-16
GA	110	3.32E-13	9.50E-15
GAQ	207	1.30E-12	1.80E-14
GAQMY	427	5.58E-12	1.30E-14

[a] Indicates the number of basis functions.

in chemical accuracy even for calculations with $> 10{,}000$ basis functions, whereas conventional methods do not yield sufficient accuracy with more than several thousand basis functions.

For reference to this partial summation technique, we consider the case of adding random numbers. We now show the algorithm for adding double-precision random numbers distributed uniformly from 0 to 1; it is a simple model of the Fock-matrix-generation algorithm:

```
C   Add random numbers
    DO k = 1, N
      TMP = 0.0
    DO l = 1, N
      RANDOM = DRAND(0)
      TMP = TMP + RANDOM
      SUM1 = SUM1 + RANDOM
    END DO
      SUM2 = SUM2 + TMP
    END DO
    PRINT SUM1, SUM2
```

where SUM1 and SUM2 denote the usual sum and partial sum, respectively. The relative numerical errors obtained by adding uniform random numbers using both methods are shown in Figure 4. The vertical line and horizontal line denote the relative numerical errors and sizes—that is, the square root of the total addition times, respectively. The size corresponds to N in the algorithm just shown. Numerical errors accumulate rapidly with the conventional method, and accuracy is improved markedly with the use of partial summation. We can analyze these results with eqs. (4) and (5). For example, with a basis size of 10,000, eq. (5) is replaced by the following equation with use of the partial summation technique:

$$(1 + 2^{-53})^{10,000} \times (1 + 2^{-53})^{10,000} - 1 \approx 2.22 \times 10^{12} \quad (6)$$

This value also exhibits fair agreement with the results of Figure 4. Thus, when the addition times are 10^{2M}, we can improve the numerical accuracy for about M digits by using the partial summation technique for every 10^{M} times (which corresponds to the size in Fig. 4), because the addition times to the same place decrease from 10^{2M} times to 10^{M} times. Surprisingly, although we observe in Figure 3 the numerical errors for total energy values, the behavior is quite similar to that in Figure 4, the simplest model of the partial summation technique. This means that the numerical accuracy for total energy values depends greatly on the numerical errors arising from the Fock matrix constructions.

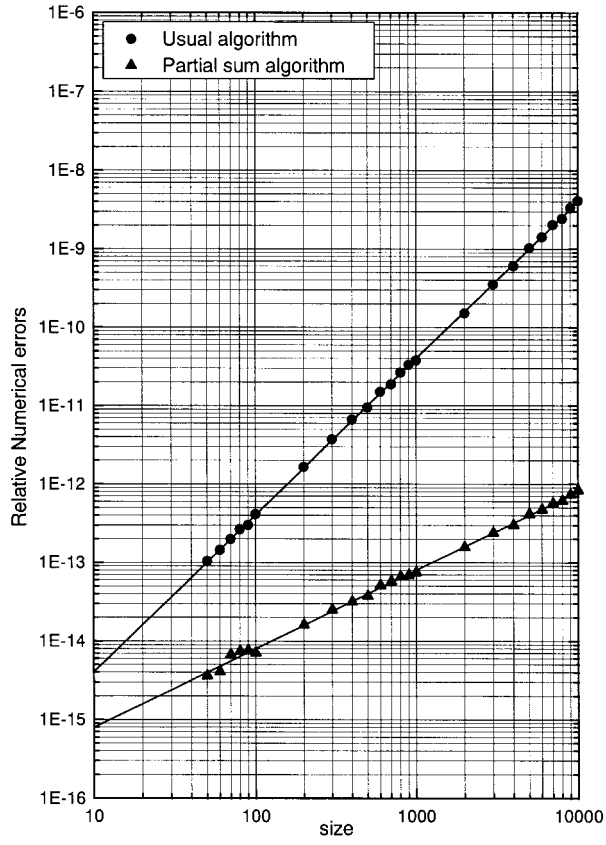

FIGURE 4. Relative numerical errors for total energy values obtained by partial summation for double-precision random numbers.

There is no mathematical or numerical reason for the identity of the slopes of the lines in Figures 3 and 4, but we have shown in this numerical experiment that the partial summation technique can markedly reduce numerical errors not only in the model case but also in the actual large-scale calculation. For execution of high-performance *ab initio* molecular orbital calculations, parallel architecture is indispensable. The results obtained show that computation of *one* Fock matrix element with parallel machines can include the partial summation technique automatically. Therefore, parallel calculation enables not only high-performance computing but also more precise numerical solutions than conventional sequential algorithms.

Another well-known summation algorithm that can dramatically decrease numerical errors is Kahan's algorithm.[19] However, the partial summation technique yields sufficient numerical accuracy

and the contrived parallel Fock matrix construction algorithm can automatically include the partial summation technique, and Kahan's method has been shown to cause a slight increase in the number of numerical computing steps. Therefore, we believe it is sufficient to use only the partial summation technique.

DIAGONALIZATION STEP

We now examine numerical accuracy in the Fock matrix diagonalization step. In the Hartree-Fock method, all eigenvalues and eigenvectors of the Fock matrix must be calculated. In most molecular orbital program packages, such as GAUSSIAN and GAMESS,[20] the Householder-QR method is usually used as the matrix diagonalization routine. We examined the numerical errors for the eigenvalues and eigenvectors obtained when we artificially introduced small numerical differences into the input matrix. For the Householder-QR method program, we used the subroutine ''hoqrvw'' created by Ninomiya[21] Notably, the same results were obtained (not shown here) when we used three other programs (e.g., the ''nshoud'' routine developed by Beppu[22]).

We tested two different kinds of matrices. One is obtained by multiplying 0.1 by the Frank matrix.[23] The matrix element, $a(i, j)$, is represented as:

$$a(i, j) = 0.1 \times \{n + 1 - \max(i, j)\}. \qquad (7)$$

We call the matrix given by eq. (7) ''Frank.'' All matrix elements are multiplied by 0.1 because the last bit of the mantissa part is always zero if we use the usual Frank matrix elements themselves. We then examined numerical errors using the matrix elements, which are multiplied by 0.1 and are given perturbation. This matrix has the merit of permitting precise estimation of numerical errors for eigenvalues, because Frank eigenvalues can be calculated analytically. The ith analytical eigenvalue, E(I), of Frank is described as follows:

```
PI = 4.0D00*DATAN(1.0D00)
E(I) = 0.1D00/(2.0D00*(1.0D00
     -DCOS(DFLOAT(2*I - 1)*PI)/
     DFLOAT(2*N + 1)))
```

A problem with Frank is that it becomes more difficult to obtain good numerical solutions in proportion to matrix size, because the pseudodegenerated eigenvalues increase. The second matrix was the real symmetry matrix whose elements are random numbers uniformly distributed from $-N$ to N, where N is the matrix size. However, the element with the maximum value in a line is exchanged for the diagonal element in this line. We call this matrix ''Random.'' Random is similar to the Fock matrix, because the diagonal elements of both matrices are larger than the offdiagonal elements.

Matrix elements are obtained by replacing the last bit of the mantissa part of the 8-byte real number by zero, and we observe the numerical accuracy by comparing the eigenvalues and eigenvectors obtained with normal results. This process is the same as in the previous section. The relative numerical error is defined as the absolute value of dividing the difference between the eigenvalues of the matrix with the last bit replaced by zero and those calculated in the usual way, with the usual eigenvalues.

Table IV shows the relative numerical errors obtained for eigenvalues and eigenvectors with various matrix sizes. The first column denotes the matrix size. The second to fourth columns show eigenvalue average errors, whereas the last two columns show eigenvector average errors. The second column shows the average relative errors of eigenvalues of Frank between the analytical solutions (Anal) and double-precision numerical solutions (Num53), $|(\text{Anal} - \text{Num53})/\text{Anal}|$ (shown as ''Anal'' in Table IV). The third column shows the average relative errors of eigenvalues of Frank between default double-precision numerical solutions and 52-bit numerical solutions with replacement of the last bit by zero (Num52), $|(\text{Num53} - \text{Num52})/\text{Num53}|$ (shown as ''Num53'' in Table IV). The fourth column shows the ''Num53'' of Random eigenvalues. The last two columns show the ''Num53'' of the Frank and Random eigenvectors, respectively. To confirm the reliability of Random, a similar analysis was performed using real symmetric matrices (RFM) appearing in the Fock matrix diagonalizations for glycine, Gly–Ala, Gly–Ala–Gln, and Gly–Ala–Gln–Met–Tyr molecules with 6-31G basis sets at the tenth iteration. RFM numerical eigenvalue errors are summarized in Table V. Figure 5 shows a logarithmic plot of the results of Random and RFM. Solid lines show the regression lines fitting the points.

For eigenvalues, the relative errors in Frank increase linearly with matrix size on the logarithmic scale. On the other hand, numerical errors in Random do not depend on matrix size or de-

TABLE IV.
Relative Numerical Errors for Eigenvalues and Eigenvectors.

	Eigenvalues			Eigenvectors	
	Frank		Random	Frank	Random
Size	Anal	Num53	Num53	Num53	Num53
64	2.81D-14	1.37D-14	4.63D-14	1.42D-12	2.40D-12
128	3.34D-14	2.75D-14	3.86D-14	3.88D-12	3.05D-12
256	1.14D-13	5.09D-14	2.78D-14	1.17D-11	8.10D-12
512	5.42D-13	1.08D-13	1.69D-14	2.10D-11	1.04D-11
1024	750D-13	3.04D-13	1.35D-14	7.81D-11	3.68D-11

creased only slightly in proportion to matrix size. This is unlike the case for Frank, possibly because the matrix characteristics in Random are better. Random has similar characteristics to the Fock matrix, and we can therefore expect results similar to those with the Fock matrix. In fact, the eigenvalue numerical errors of RFM were smaller than those of Random, and the behavior (i.e., size dependency) was very similar to that of Random. Although the number of samples tested was very small, we can expect that the average numerical error of Fock matrix eigenvalues would be smaller than the required numerical accuracy even if the matrix size = 10,000. The eigenvalue with the maximal numerical error usually corresponds to the highest virtual orbital energy, because global origin shift is usually applied in the Fock matrix diagonalization. Eigenvalues corresponding to occupied orbitals appear to have numerical errors of about 1.0D-13.

We next consider the numerical errors of the eigenvectors. Relative errors of eigenvectors are larger than those of eigenvalues. However, errors in Random and RFM, depicted in Figure 5, increase with matrix size, but the estimated error is a maximum of about 1.0D-10 by extrapolation, even for 10,000 basis functions. These values are within the required chemical accuracy in both cases.

Consequently, the behaviors of numerical errors of eigenvalues and eigenvectors depend greatly on the characteristics of the matrix. However, numerical accuracy of Random eigenvalues does not depend on matrix size but is almost constant across matrix size, whereas for Random eigenvectors it depends on matrix size. The average numerical error is expected to be between 1.0D-14 and 1.0D-10. Because the characteristics of Random appear to be quite similar to those of the Fock matrix, it can be stated that the numerical errors in the Fock

TABLE V.
Relative Numerical Errors for RFM Eigenvalues.

Molecule	Size	Average
Gly	55	5.611D-15
GA	110	9.317E-15
GAQ	207	8.119D-15
GAQMY	427	9.099E-15

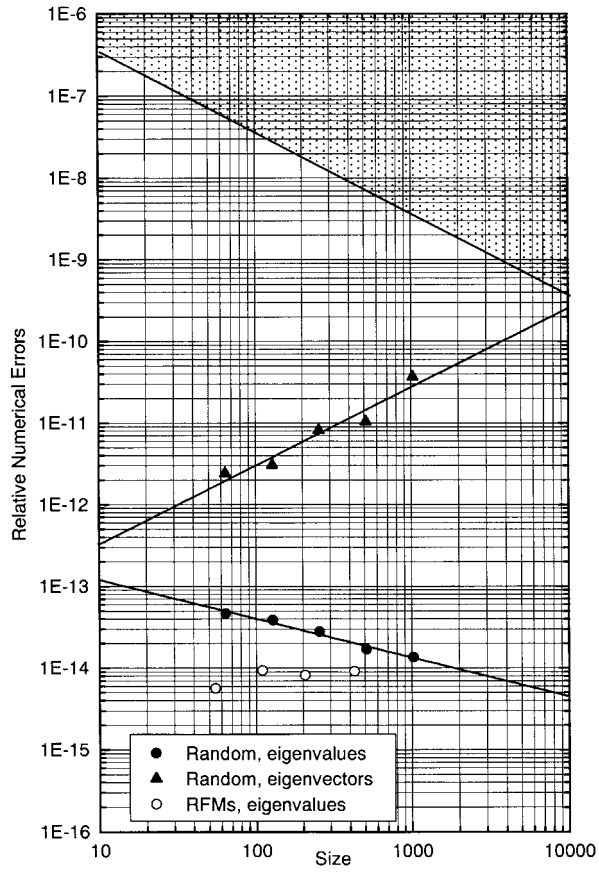

FIGURE 5. Average numerical computing errors and required accuracy for eigenvalues and eigenvectors.

matrix diagonalization step are within the required chemical accuracy if the matrix size is not $> 10{,}000$.

Conclusions

The partial summation technique in the Fock matrix generation step can reduce the numerical error and keep chemical accuracy for total energy values of molecules even if basis size is $> 10{,}000$, whereas the conventional method makes for a numerical error larger than chemical accuracy. On the other hand, the numerical error of the Householder-QR diagonalization routine is equal to or less than chemical accuracy, even with the matrix size of 10,000. Computation of *one* Fock matrix element with parallel machines can include the partial summation technique automatically, so parallel calculation yields not only high-performance computing but also more precise numerical solutions than the conventional sequential algorithm. Computing systems and algorithms for large-scale calculations require designs suitable to problem size and required accuracy.

Acknowledgments

We thank Dr. S. Obara, Dr. T. Amisaki, Dr. K. Murakami, Mr. S. Inabata, Mr. S. Yamada, Mr. N. Miyakawa, and Dr. O. Kitao, for much helpful advice and discussion. We are also very grateful to the referees for suggestions and comments.

References

1. Cioslowski, J. In Reviews in Computational Chemistry IV; Lipkowitz, K. B.; Boyd, D. B., Eds., VCH: New York, 1993; p 1.
2. Mattson, T. G. In Parallel Computing in Computational Chemistry ACS Symposium Series; American Chemical Society: Washington, DC, 1995.
3. Brode, S.; Horn, H.; Ehrig, M.; Moldrup, D.; Rice, J. E.; Ahlrichs, R. J Comput Chem 1993, 14, 1142.
4. Challacombe, M.; Schwegler, E. J Chem Phys 1997, 106, 5526.
5. Strout, D. L.; Scuseria, G. J Chem Phys 1995, 102, 8448.
6. Furlani, T. R.; King, H. F. J Comput Chem 1995, 16, 91.
7. Foster, I. T.; Tilson, J. L.; Wagner, A. F.; Shepard, R. L.; Harrison, R. J.; Kendall, R. A.; Littlefield, R. J. J Comput Chem 1996, 17, 109.
8. Strain, M. C.; Scuseria, G. E.; Frisch, M. J. Science 1996, 271, 51.
9. Schwegler, E.; Challacombe, M. J Chem Phys 1996, 105, 2726.
10. Nagashima, U.; Obara, S.; Murakami, K.; Yoshii, T.; Shirakawa, S.; Amisaki, T.; Kitamura, K.; Takashima, H.; Tanabe, K. IPSJ SIGNotes 1996, 117-ARC-16, 89.
11. Nagashima, U.; Obara, S.; Murakami, K.; Amisaki, T.; Kitamura, K.; Takashima, H.; Kitao, O.; Tanabe, K.; Inabata, S.; Yamada, S.; Miyakawa, N. JCPE Newsletter 1997, 9, 1.
12. Machine epsilon is the smallest floating-point number which, when added to the floating-point number 1.0, produces a floating-point result different from 1.0, and is also called machine accuracy. We note that it is not the floating-point number that can be represented on a machine. In SGI-R10000 (IEEE standard), machine epsilon is 2.22D-16. See, for example: Press, W. H.; Teukolski, S. A.; Vetterling, W. T.; Flannery, B. P. In Numerical Recipes in C, 2nd ed.; Cambridge University Press: New York, 1992; p 28, 6889.
13. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. GAUSSIAN-94, Revision D.4, Gaussian, Inc., Pittsburgh, PA, 1995.
14. Kudo, T.; Hashimoto, F.; Gordon, M. S. J Comput Chem 1996, 17, 1163.
15. Cioslowski, J.; O'Connor, P. B.; Fleischmann, E. D. J Am Chem Soc 1991, 113, 1086.
16. Cioslowski, J. J Am Chem Soc 1991, 113, 4139.
17. Takaoka, Y. Private communication.
18. Sakuma, T.; Kashiwagi, H.; Takada, T.; Nakamura, H. Int J Quantum Chem 1997, 61, 137.
19. Kahan, W. Commun ACM 1965, 8, 40.
20. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. J Comput Chem 1993, 14, 1347.
21. Ninomiya, I. NUMPACK, program library of the Computer Center of Nagoya University.
22. Beppu, Y. NICER, program library of the Computer Center of Nagoya University.
23. Frank, W. L. J Soc Indust Appl Math 1958, 6, 378.